

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-003361

(43)Date of publication of application : 07.01.2000

(51)Int.Cl.

G06F 17/27  
G06F 17/30

(21)Application number : 10-168055

(71)Applicant : DAINIPPON PRINTING CO LTD

(22)Date of filing : 16.06.1998

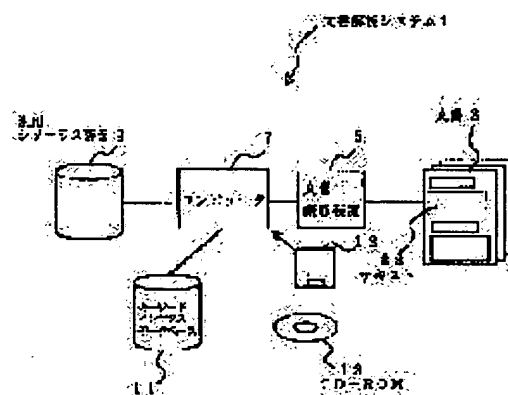
(72)Inventor : FUJIOKA TAKAKO

## (54) DOCUMENT ANALYSIS SYSTEM, AND RECORDING MEDIUM

### (57)Abstract:

**PROBLEM TO BE SOLVED:** To provide a document analysis system which analyzes a document in a paper medium publication and holds information on its contents as a data base.

**SOLUTION:** A computer 7 is provided with a document reader 5, a general purpose thesaurus dictionary 9, and a keyword index data base 11. The document reader 5 reads a text of a document 3 printed on a paper medium. Also, when printing data of the document 3 exist, the computer 7 may be made to input them. The general purpose thesaurus dictionary 9 classifies every word into a conceptual hierarchy type and registers it. The keyword index data base 11 registers a word appearing in the document analyzed by the computer 7 with an index attached. A CD-ROM 13 records a program by which the computer 7 analyzes the document, classifies and registers the word.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

**\* NOTICES \***

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. \*\*\*\* shows the word which can not be translated.
3. In the drawings, any words are not translated.

---

**CLAIMS**

---

**[Claim(s)]**

[Claim 1] The document analysis system characterized by providing the means which is a document analysis system for analyzing a document, collates a thesaurus dictionary, and the word and said thesaurus dictionary in said document, and takes out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and a maintenance means to hold the taken-out word.

[Claim 2] The document analysis system characterized by to provide a maintenance means hold the means which is a document analysis system for analyzing a document, collates a thesaurus dictionary, and the word and said thesaurus dictionary in said document, and takes out the superordinate-concept item of the word concerned, and the word concerned about a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and the taken-out superordinate-concept item and a word.

[Claim 3] Said maintenance means is claim 1 characterized by attaching and holding an index in the taken-out word, or a document analysis system according to claim 2.

[Claim 4] The record medium which recorded the program for making it function as collating the word and thesaurus dictionary in a document for a computer, taking out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and holding the taken-out word.

[Claim 5] The record medium which recorded the program for making it function as collating the word and thesaurus dictionary in a document for a computer, taking out the superordinate-concept item of the word concerned, and the word concerned about a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and holding the superordinate-concept item and word which were taken out.

---

[Translation done.]

\* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.\*\*\*\* shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

---

DETAILED DESCRIPTION

---

[Detailed Description of the Invention]

[0001]

[Field of the Invention] This invention relates to a document analysis system and a record medium.

[0002]

[Description of the Prior Art] Treating an electronic filing document in everyday life these days by the spread of personal computers etc. is increasing. An electronic filing document has the format which divides into the appearance information which shows the appearance word information which shows the contents of the document, and the logical structure, and is held and which was standardized [ SGML ], and maintenance, a reorganization collection, etc. of document data are performed comparatively freely. Although an electronic filing document is increasing, in the site of actual publication and printing, in large quantities, the conventional paper medium publications, such as an informational magazine and a catalog, are produced, and are circulating.

[0003]

[Problem(s) to be Solved by the Invention] However, since the document information printed by such paper medium publication is printed for various appearance and is immediately discarded even if the text section is created as electronic data in the process before printing, it does not have a means to hold as document data like an electronic filing document, and is not easy to reuse.

[0004] The place which this invention was made in view of such a problem, and is made into the purpose analyzes the document in a paper medium publication, and is to offer the document analysis system which holds the information about the contents as a database, and the record medium as which a computer is operated as a document analysis system.

[0005]

[Means for Solving the Problem] It is the document analysis system characterized by providing the means which this invention is a document analysis system for analyzing a document, collates a thesaurus dictionary, and the word and said thesaurus dictionary in said document, and takes out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and a maintenance means to hold the taken-out word in order to attain the purpose mentioned above.

[0006] Moreover, the 2nd invention is the record medium which recorded the program for making it function as collating the word and thesaurus dictionary in a document for a computer, taking out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and holding the taken-out word.

[0007]

[Embodiment of the Invention] Hereafter, the gestalt of operation of this invention is explained to a detail based on a drawing. Drawing 1 is drawing showing the outline configuration of the document analysis system 1 concerning the gestalt of this operation.

[0008] In drawing 1 , the document reader 5, the general-purpose thesaurus dictionary 9, and the keyword index database 11 are formed in a computer 7. The document reader 5 reads the text 33 of the document 3 printed by the paper medium, and is OCR etc. moreover, the data for printing of a document 3 -- the inside of a floppy disk 12 -- text data -- \*\* -- it carries out, and when it exists, this can also be made to input into a computer 7 The general-purpose thesaurus dictionary 9 classifies and registers all words into a notional hierarchical type. In addition, the general-purpose thesaurus dictionary 9 may be held at CD-ROM etc.

[0009] The word which appears in the document analyzed by computer 7 can attach an index to the keyword index database 11, and is registered into it. The detail of this classification and registration processing is explained later. A computer 7 analyzes a document in CD-ROM13, and the program for classifying and registering a word is recorded on

it.

[0010] Drawing 2 is a flow chart which shows analysis processing of the document 3 by the document analysis system 1, and drawing 3 is drawing showing an example of the general-purpose thesaurus dictionary 9 and the keyword index database 11. The document analysis system 1 reads the text 33 of a document 3, disassembles a text into a noun, a verb, etc. (step 201), and extracts the noun in a document (step 202). At this time, the computer 7 shown in drawing 1 has a dictionary for language analysis etc. inside, and extracts the noun in a document with reference to them.

[0011] Next, it refers to the word in the general-purpose thesaurus dictionary 9, and the count of document Nakade present of lower \*\*\*\* is computed for every superordinate concept (step 203). When it is judged and (step 204) exceeded whether the count of an appearance of lower \*\*\*\* of a certain superordinate-concept node exceeds a threshold, an index 311 is attached to the superordinate-concept node 303 concerned and the lower group word group 305, and it registers with the keyword index database 11 (step 205).

[0012] Moreover, when the count of an appearance of lower \*\*\*\* does not exceed a threshold at step 204, the superordinate-concept node 303 concerned and the lower group word group 305 are not registered into abandonment (step 206) 11, i.e., a keyword index database.

[0013] The document 3 shown in drawing 3 presupposes that it is a document about cooking. The document reader 5 shown in drawing 1 reads a document 3, and a computer 7 extracts the noun in a document. Furthermore, a computer 7 collates the extracted noun and the word registered into the general-purpose thesaurus dictionary 9, and computes the count of an appearance in a document.

[0014] Here, the general-purpose thesaurus dictionary 9 registers the word of various fields according to a concept hierarchical, as shown in drawing 3. The superordinate-concept node 303 subdivided still more notionally is registered into the general-purpose thesaurus dictionary 9 to the superordinate concept 301 without lower \*\*\*\*, and the lower group word group 305 which belongs to the superordinate-concept node 303 of them notionally is registered into it.

[0015] Supposing a document 3 is related with cooking, many words about the food containing an ingredient etc. are contained in the noun in a document 3. Therefore, if the count of an appearance in the document 3 of each \*\*\*\*\* is computed with reference to the general-purpose thesaurus dictionary 9, counts of an appearance, such as lower group word group 305b belonging to lower group word group 305a to which superordinate-concept node 303 in general-purpose thesaurus dictionary 9 a "a seasoning" belongs, or superordinate-concept node 303b "vegetables", will increase.

[0016] If the threshold in step 204 is made into "30 times" here, an index 331 will be attached to each \*\*\*\*\* and superordinate-concept node 303a whose count of an appearance is 35 times as shown in drawing 3, its lower group word group 305a, and superordinate-concept node 303b and its lower group word group 305b will be registered into the keyword index database 11.

[0017] In addition, the index 331 shown in drawing 3 is attached also to the noun which appears in a document 3, holds and manages the text data of a document 3, and when searching or reusing the text 33 of a document 3 behind, it is used. Moreover, if a document 3 is limited to that to which the contents are similar, registration processing of the appearance word to the keyword index database 11 can carry out more efficiently.

[0018] Superordinate-concept node 303c with which the count of an appearance of lower group word group 305c shown in drawing 3 did not fill a threshold on the other hand, and its lower group word group 305c are not registered into the keyword index database 11. That is, it is supposed that the contents of the document 3 are this lower group word group 305c that unrelated [ almost ].

[0019] Thus, according to the gestalt of operation of this invention, the word which appears frequently in the document 3 printed by the paper medium is efficiently registered into the keyword index database 11. The keyword index database 11 is used in case the contents of the document 3 are used as the document data of SGML correspondence. In addition, the threshold of the count of an appearance is determined by the amount and the contents of the document 3 made to read into a computer 7.

[0020]

[Effect of the Invention] As mentioned above, as explained to the detail, according to this invention, the contents of the document printed by the paper medium are data-ized efficiently, and it can use for next-izing corresponding to an SGML document.

---

[Translation done.]

\* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. \*\*\*\* shows the word which can not be translated.
3. In the drawings, any words are not translated.

---

TECHNICAL FIELD

---

[Field of the Invention] This invention relates to a document analysis system and a record medium.

---

[Translation done.]

**\* NOTICES \***

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.\*\*\*\* shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

---

**PRIOR ART**

---

[Description of the Prior Art] Treating an electronic filing document in everyday life these days by the spread of personal computers etc. is increasing. An electronic filing document has the format which divides into the appearance information which shows the appearance word information which shows the contents of the document, and the logical structure, and is held and which was standardized [ SGML ], and maintenance, a reorganization collection, etc. of document data are performed comparatively freely. Although an electronic filing document is increasing, in the site of actual publication and printing, in large quantities, the conventional paper medium publications, such as an informational magazine and a catalog, are produced, and are circulating.

---

[Translation done.]

\* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.\*\*\*\* shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

---

EFFECT OF THE INVENTION

---

[Effect of the Invention] As mentioned above, as explained to the detail, according to this invention, the contents of the document printed by the paper medium are data-ized efficiently, and it can use for next-izing corresponding to an SGML document.

---

[Translation done.]



**\* NOTICES \***

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. \*\*\*\* shows the word which can not be translated.
3. In the drawings, any words are not translated.

---

**TECHNICAL PROBLEM**

---

[Problem(s) to be Solved by the Invention] However, since the document information printed by such paper medium publication is printed for various appearance and is immediately discarded even if the text section is created as electronic data in the process before printing, it does not have a means to hold as document data like an electronic filing document, and is not easy to reuse.

[0004] The place which this invention was made in view of such a problem, and is made into the purpose analyzes the document in a paper medium publication, and is to offer the document analysis system which holds the information about the contents as a database, and the record medium as which a computer is operated as a document analysis system.

---

[Translation done.]

\* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.\*\*\*\* shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

---

MEANS

---

[Means for Solving the Problem] It is the document analysis system characterized by providing the means which this invention is a document analysis system for analyzing a document, collates a thesaurus dictionary, and the word and said thesaurus dictionary in said document, and takes out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and a maintenance means to hold the taken-out word in order to attain the purpose mentioned above.

[0006] Moreover, the 2nd invention is the record medium which recorded the program for making it function as collating the word and thesaurus dictionary in a document for a computer, taking out a word with the high frequency of occurrence within said thesaurus dictionary with the word in said document, and holding the taken-out word.

[0007]

[Embodiment of the Invention] Hereafter, the gestalt of operation of this invention is explained to a detail based on a drawing. Drawing 1 is drawing showing the outline configuration of the document analysis system 1 concerning the gestalt of this operation.

[0008] In drawing 1, the document reader 5, the general-purpose thesaurus dictionary 9, and the keyword index database 11 are formed in a computer 7. The document reader 5 reads the text 33 of the document 3 printed by the paper medium, and is OCR etc. moreover, the data for printing of a document 3 -- the inside of a floppy disk 12 -- text data -- \*\* -- it carries out, and when it exists, this can also be made to input into a computer 7 The general-purpose thesaurus dictionary 9 classifies and registers all words into a notional hierarchical type. In addition, the general-purpose thesaurus dictionary 9 may be held at CD-ROM etc.

[0009] The word which appears in the document analyzed by computer 7 can attach an index to the keyword index database 11, and is registered into it. The detail of this classification and registration processing is explained later. A computer 7 analyzes a document in CD-ROM13, and the program for classifying and registering a word is recorded on it.

[0010] Drawing 2 is a flow chart which shows analysis processing of the document 3 by the document analysis system 1, and drawing 3 is drawing showing an example of the general-purpose thesaurus dictionary 9 and the keyword index database 11. The document analysis system 1 reads the text 33 of a document 3, disassembles a text into a noun, a verb, etc. (step 201), and extracts the noun in a document (step 202). At this time, the computer 7 shown in drawing 1 has a dictionary for language analysis etc. inside, and extracts the noun in a document with reference to them.

[0011] Next, it refers to the word in the general-purpose thesaurus dictionary 9, and the count of document Nakade present of lower \*\*\*\* is computed for every superordinate concept (step 203). When it is judged and (step 204) exceeded whether the count of an appearance of lower \*\*\*\* of a certain superordinate-concept node exceeds a threshold, an index 311 is attached to the superordinate-concept node 303 concerned and the lower group word group 305, and it registers with the keyword index database 11 (step 205).

[0012] Moreover, when the count of an appearance of lower \*\*\*\* does not exceed a threshold at step 204, the superordinate-concept node 303 concerned and the lower group word group 305 are not registered into abandonment (step 206) 11, i.e., a keyword index database.

[0013] The document 3 shown in drawing 3 presupposes that it is a document about cooking. The document reader 5 shown in drawing 1 reads a document 3, and a computer 7 extracts the noun in a document. Furthermore, a computer 7 collates the extracted noun and the word registered into the general-purpose thesaurus dictionary 9, and computes the count of an appearance in a document.

[0014] Here, the general-purpose thesaurus dictionary 9 registers the word of various fields according to a concept hierarchical, as shown in drawing 3. The superordinate-concept node 303 subdivided still more notionally is registered into the general-purpose thesaurus dictionary 9 to the superordinate concept 301 without lower \*\*\*\*, and the lower

group word group 305 which belongs to the superordinate-concept node 303 of them notionally is registered into it.

[0015] Supposing a document 3 is related with cooking, many words about the food containing an ingredient etc. are contained in the noun in a document 3. Therefore, if the count of an appearance in the document 3 of each \*\*\*\*\* is computed with reference to the general-purpose thesaurus dictionary 9, counts of an appearance, such as lower group word group 305b belonging to lower group word group 305a to which superordinate-concept node 303 in general-purpose thesaurus dictionary 9 a "a seasoning" belongs, or superordinate-concept node 303b "vegetables", will increase.

[0016] If the threshold in step 204 is made into "30 times" here <A To HREF="/Tokujitu/tjitemdrw.ipdl?N0000=237&N0500=1E\_N/;>=?? <<9>///&N0001=505&N0552=9&N0553=000005" TARGET="tjitemdrw"> drawing 3 An index 331 is attached to each \*\*\*\*\* and superordinate-concept node 303a whose count of an appearance is 35 times so that it may be shown, its lower group word group 305a, and superordinate-concept node 303b and its lower group word group 305b are registered into the keyword index database 11.

[0017] In addition, the index 331 shown in drawing 3 is attached also to the noun which appears in a document 3, holds and manages the text data of a document 3, and when searching or reusing the text 33 of a document 3 behind, it is used. Moreover, if a document 3 is limited to that to which the contents are similar, registration processing of the appearance word to the keyword index database 11 can carry out more efficiently.

[0018] Superordinate-concept node 303c with which the count of an appearance of lower group word group 305c shown in drawing 3 did not fill a threshold on the other hand, and its lower group word group 305c are not registered into the keyword index database 11. That is, it is supposed that the contents of the document 3 are this lower group word group 305c that unrelated [ almost ].

[0019] Thus, according to the gestalt of operation of this invention, the word which appears frequently in the document 3 printed by the paper medium is efficiently registered into the keyword index database 11. The keyword index database 11 is used in case the contents of the document 3 are used as the document data of SGML correspondence. In addition, the threshold of the count of an appearance is determined by the amount and the contents of the document 3 made to read into a computer 7.

---

[Translation done.]

\* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. \*\*\*\* shows the word which can not be translated.
3. In the drawings, any words are not translated.

---

DESCRIPTION OF DRAWINGS

---

[Brief Description of the Drawings]

[Drawing 1] Drawing showing the document analysis system 1 concerning the gestalt of 1 operation of this invention

[Drawing 2] The flow chart which shows the analysis processing by the document analysis system 1

[Drawing 3] Drawing showing the general-purpose thesaurus dictionary 9 and the keyword index database 11

[Description of Notations]

1 ..... Document analysis system

3 ..... Document

5 ..... Document reader

7 ..... Computer

9 ..... General-purpose thesaurus dictionary

11 ..... Keyword index database

---

[Translation done.]

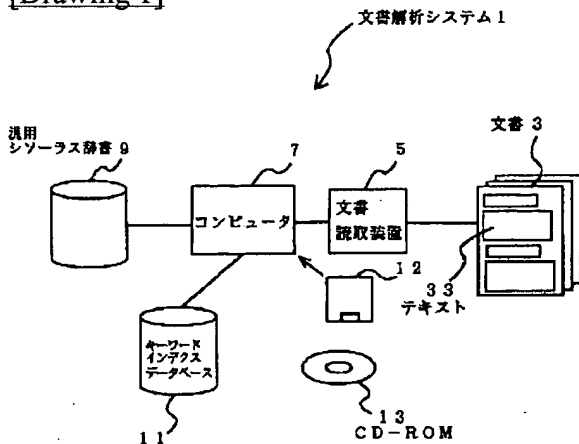
## \* NOTICES \*

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

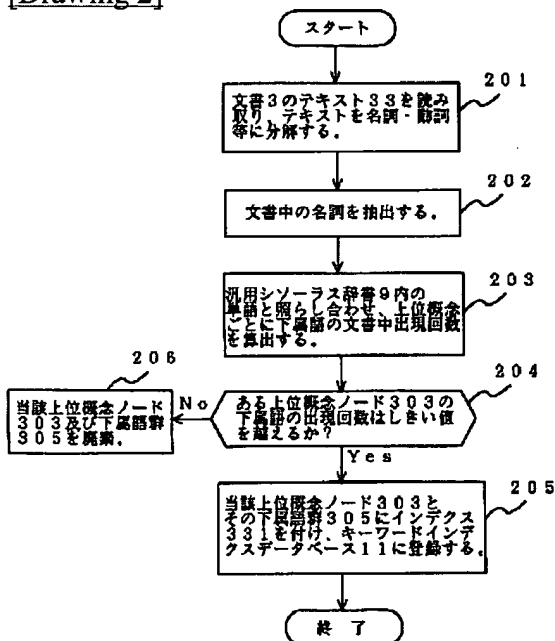
1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. \*\*\*\* shows the word which can not be translated.
3. In the drawings, any words are not translated.

## DRAWINGS

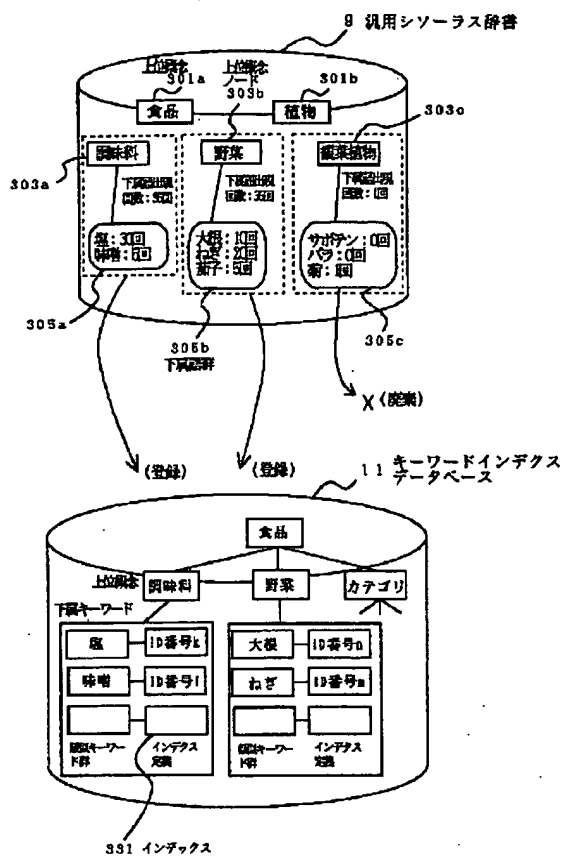
[Drawing 1]



[Drawing 2]



[Drawing 3]



[Translation done.]